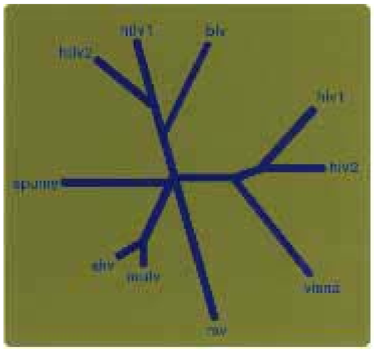
Bioinformatics at the ABCC

A Research Project of Gary Smythers and Dr. Robert M. Stephens

Resources

The ABCC is known as an outstanding bioinformatics resource for molecular biologists, providing computational tools for a complete range of biological research problems, from primary sequence determinate to molecular modeling and gene linkage analysis. Numerous application program and packages are available for performing these diverse tasks. Additional programs provide an interface between sequence data and structural databases including many modeling and visualization packages. Together, these programs represent a dynamic analysis environment that is constantly being improved with novel and improved algorithms to accommodate user requests and aid in analysis. In addition to the analysis software, the ABCC maintains a complete set of current molecular biological databases including daily updates of genbank. Genbank now has 1.4 billion bases and is nearly doubling in size yearly. To further compound the increase in size of the genetic databases, the annotation of the sequences represents almost twice as much space as the sequences themselves.

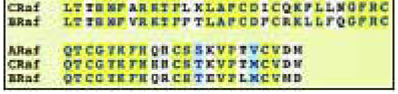


F-1. Phylogenic Tree for Retroviral Reverse Transcriptases.

The ABCC also generates two databases for local use and for distribution to other sites via our anonymous ftp facility. These are a database containing the translated products of genbank, and non-redundant protein and nucleic acid databases used for similarity searching on various platforms including the massively parallel Maspar. The non-redundant database is derived from all published nucleotide data, including "tag" EST sequences. Databases of structural data include coordinates for small molecules and proteins as well as nucleic acids. The dramatic growth of these databases means that the ABCC must address increasingly complex user queries which require additional CPU power and data storage. As comparative and experimental information about the sequences grows, this additional information will need to be incorporated into additional databases and require the development of new analysis tools. In order to perform bioinformatics analysis, the ABCC houses a heterogenous mix of interconnected platforms form workstations to superscalar and supervector computers. Such a diverse mix allows applications to be configured for the most appropriate machine in terms of utility and efficiency. This heterogenous array is well integrated and NQS queueing is used to allow distribution of requests across this network. Distributed processing allows the ABCC to more fully utilize CPU cycles to meet the increasing demands form the research community. Since we anticipate tremendous increases in both the number and complexity of these analyses, the existing hardware will eventually be outstripped by the user demand. One of the advantages to scientists using the ABCC is the personalizes consultation services that are available. The bioinformatics support staff has experience in both the biological and computer sciences and can provide a beneficial link between the scientist and the computational tools needed in their research. This laboratory experience facilitates the translation of research queries from laboratory scientists into results of complex analyses performed by the support staff. Frequently, users present specific requests that can only be satisfied by modifications of existing software or the development of new programs. The ABCC can then apply these additional analysis tools to other requests as they are generated to provide an ever-increasing resource set. The addition of further bioinformatics personnel with diverse research experiences and backgrounds enrolled in software development will insure that the ABCC continues to provide this good relationship with the scientific community.

WEB and Network Access

One of the important approaches to the software development and implementation at the ABCC is to make the analysis programs as consistent and intuitive as possible for the users. The goal of this approach is to provide maximal usage to even our less experienced scientists so that they can forgo extensive computer training or preparation. One of the biggest aspects of this has been the development of extensive enhancements to data visualization, allowing for facilitated interpretation of results. In order to further increase the utilization of our bioinformatics products, we have begun to develop WEB and GUI (Graphics user interface) based access to these programs.



```
ARaf   LTNHNFVRKTFFSLAFCDFCLKFLFNGFRC
CRaf   LTHNNFARKTFLKLAFCDICQKFLLNGFRC
BRaf   LTTHNFVRKTFFTLAFCDFCRKLLFQGFRC

ARaf   QTCGYKFHQHCSSKVPTVCVDM
CRaf   QTCGYKFHEHCSTKVPTMCVDW
BRaf   QTCGYKFHQRCSTEVPLMCVND
```

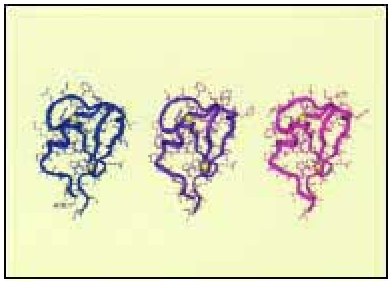
F-2a. Alignment of Cysteine-Rich Regions for Raf Family.

The completion of the many genome projects currently underway represents only the first step in the analysis of the huge volumes of data being produced. Manipulation of larger and larger DNA sequences and sets of DNA sequences will continue to challenge existing computer resources at the level of the cpu, memory and storage capacity. There are two important areas where the volume of information being generated will required increased computing resources. First, the accommodation of these longer sequences will require modification of the existing sequence analysis software. Also, with data production becoming more and more automated, the ability to make comparisons between the thousands of newly identified sequences and the entire existing databases will become more time consuming. Thus, the sequences being analyzed are becoming more complex and the number of sequences that must be compared is increasing. Finally, the nature of the full genome sequences will require the development of new comparison algorithms and software which has had no previous parallel.

Sequence Determination and Phylogenetic Analysis

As the number of sequences and sources of sequences has increased, it has allowed to relatedness of proteins across relatively long evolutionary periods to be estimated. As expected, this has necessitated the use of increased computing resources and development of more robust software tools. As an example of the utility of these tools, consider the retroviral reverse transcriptases. The polymerases perform replication of the viral genome and are thus important pharmacological intervention targets. In the figure to the right, we have aligned the sequences from several related groups of retroviruses and displayed the results in a phylogenetic tree.

Sequence Analysis and Structure Prediction



F-2b. Structure Prediction for A-Raf (purple) and B-raf (magneta) Using c-Raf (blue).

As the databases of both sequence and structural information have expanded, the ability to use existing structural information to gain insights about possible structures of molecules for which there is no structural information have evolved considerably. Using a homologous structure, or better yet, a family of related structures, it has become possible to predict some structural information about a new sequence. Because of the long lag time between the generation of the sequence for a potential protein and the determination of structural information for a protein, the need for this type of analysis will be constantly increasing. These first structural clues may lead towards identifying potential drugs or other molecules that may interact with a new protein, or alternatively, the identification of residues that might be good targets for mutagenesis studies involving the protein. One of the principal needs in this type of analysis is for high-speed graphics workstations that can access compute servers where the compute intensive portions of the analysis are performed. Furthermore, as the methods of prediction evolve, they will certainly require increased computer resources because their complexity will increase and because the number of analyses being performed will increase. As an example of the application of this type of analysis, consider the Raf family of serine/threonine kinases. These related molecules all contain a cysteine finger region, that in c-Raf binds phosphatidyl-serine and 14-3-3 proteins. The structure of the c-Raf cysteine finger has been determined (1). In addition, the structure of a related structure has been determined for the protein kinase C proteins (2). As these molecules are differentially regulated, they might bind to different lipids in this crucial activation site. In Figure F-2z, the alignment of the cysteine finger regions of the three Raf molecules is presented. The the second panel (Figure F-2b), cartoons of the structure for the c-Raf region and the predicted structures for the other two family members (A-Raf and B-Raf) are shown.